PeritumAI

WHITE PAPER 05

AI THREATS // WHITE PAPER

The Attacker Has the Same Models

Defending at machine speed against
AI-enabled attacks, jailbreaks and deepfakes

PeritumAI

AI THREATS // WHITE PAPER 05

The Attacker Has the Same Models

Defending at machine speed against AI-enabled attacks, jailbreaks and deepfakes

Published
25 October 2026

Edition
Version 1.1

Evidence window
17 Sep to 22 Oct 2026

> **How to read this paper.** Every incident below comes from public reporting. Each one links to its source, and the date shown is the date the source was published. We checked every fact against the archived text of the source on 25 October 2026.
>
> This is general education, not legal advice. Rules change and specifics vary. Check with a qualified professional for your situation. Current as of 25 October 2026.

Executive summary

The AI models your business uses to write code and summarise documents are now part of the attacker's toolkit too. In 2026 that moved from forecast to incident record.

In September and October 2026, researchers and incident responders documented:

- malware that asks four commercial AI models to vote on its next move;[2]
- a criminal group using an AI coding agent to run intrusions that harvested more than 628,000 records;[10]
- a China state-aligned actor using an AI penetration-testing framework to try to sign in to more than 2,250 Microsoft 365 tenants, at up to 2,900 attempts a minute;[17]
- an AI-driven attack that went from a foothold in a helpdesk system to root "in seconds".[5]

On the research side, an XBOW researcher, directing AI agents, built a full Google Chrome exploit chain in six days with a publicly available AI model.[12] Another reproduced previously disclosed refrigeration-controller vulnerabilities, and a working root exploit, in about 28 minutes for about $6 of AI usage.[11]

AI does not make attackers unstoppable. Recorded Future noted that AI-assisted attackers still make basic operational mistakes.[14] What AI changes is time. The gap between a weakness existing and a weakness being exploited is shrinking. Defences that count on attackers being slow, such as monthly patching, manual triage and once-a-year phishing training, are losing their margin.

This paper reviews the evidence, explains what changes for defenders, and sets out a control framework mapped to ISO/IEC 42001, the NIST AI RMF, the EU AI Act, the OWASP Top 10 for LLM Applications and the GDPR. It covers three linked threats: AI-accelerated intrusion, jailbroken and abused models, and AI-generated deception.

The problem

Speed is the new gap between attacker and defender

Turning a vulnerability into a working exploit, and an exploit into a campaign, used to take skilled people and time. AI agents compress both. In Booz Allen's operational technology lab, AI models went from a perimeter compromise to actions inside an industrial control network in just over 16 minutes, and met their objectives in all eight test scenarios.

The testers made the agents wait for human approval before exploiting anything or causing physical effects, so real attackers would face fewer brakes.[9] DIVD described the agent that attacked it as working "at the speed of light and sloppy logic or pattern".[5]

Your AI keys can pay for attacks on others

Criminal tools now run on stolen AI provider keys. Bitsight tracked MONST3R-AI, sold as an autonomous hacking tool that runs scans on stolen OpenAI, Anthropic and Google keys, so the key's owner pays.[16] SecurityWeek reported a botnet, x47.c, with an "AI API drain" mode to burn through a victim's paid AI credits.[3] A leaked key is now a financial and a reputational liability.

Guardrails slow attackers down. They do not stop them.

Jailbreaking means getting a model to ignore its safety rules. It remained the biggest topic on the underground forums Bitsight tracked, and talk about it grew by roughly three quarters between February and August 2026.[16] Research keeps finding new angles. UNSW Sydney researchers showed that models prompted or fine-tuned to write like drunk people became easier to jailbreak and more likely to leak secrets shared in confidence.[4] On criminal forums, members run open-weight models where no provider checks requests, and some strip out the safety behaviour themselves.[16] When a model refused, the CyberXero actor often reopened a fresh session with an "authorized pentest" framing preloaded and tried again, though some refusals held.[10]

Deception has become cheap

Generative AI makes convincing fake content cheap to produce. The Manhattan District Attorney seized 12 deepfake websites hosting AI-generated videos of more than 1,200 people.[1] Business-facing deception, such as cloned executive voices, fake documents and tailored lures, uses the same technology.

Incident evidence

These incidents come from PeritumAI's daily incident-intelligence monitoring. Dates are publication dates of the cited source.

Date	What happened	What it shows	Source
17 Sep 2026	**Deepfake site seizure.** The Manhattan District Attorney seized 12 deepfake websites hosting AI-generated videos of more than 1,200 people.	Synthetic media at industrial scale	[1]
24 Sep 2026	**CLOSEDQUORUM.** Cisco Talos documented Windows malware that hands its tactical decisions to four commercial LLMs (DeepSeek, Qwen, Mistral and Google Gemini), counts their votes and acts on the majority, aiming to steal credentials and crypto wallets. No human operator or dedicated command server is needed. The public sample Talos analysed was inert, with placeholder keys.	AI as malware command and control	[2]
28 Sep 2026	**"Drunk" models leak more.** UNSW Sydney researchers found that models prompted or fine-tuned to write like drunk people became easier to jailbreak and more likely to reveal confidential information.	Model safety is fragile under prompting and fine-tuning	[4]
1 Oct 2026	**Agentic intrusion at DIVD.** On 21 September an AI-driven attack chained two Zammad zero-days and escalated from Zammad's local system user to root in seconds, then read and exfiltrated data.	Agentic attacks compress the attack chain	[5]

Date	What happened	What it shows	Source
1 Oct 2026	**AI policy experts targeted.** Proofpoint reported a China-aligned group, TA419, impersonating AI policy figures, including a senior Anthropic employee and a former White House official, to phish AI policy experts.	The AI sector itself is a target	[6]
5 Oct 2026	**An AI-found bug is being probed.** A Rejetto HFS flaw (CVE-2026-61500, CVSS 9.3), found by Horizon3.ai with help from Anthropic's Mythos model, is now being probed by attackers, according to VulnCheck. Activity so far looks like small-scale reconnaissance.	AI-assisted discovery feeds the exploitation pipeline	[8]
5 Oct 2026	**Google paused open-source bug bounty submissions.** Google stopped accepting product vulnerability reports to its open-source rewards programme because of a sharp rise in automated submissions, most of them invalid.	Automated noise overwhelms defender processes	[7]
11 Oct 2026	**AI attacks on operational technology.** In Booz Allen lab tests, AI models met their objectives in all eight industrial attack scenarios, in one case finding and moving a robotic arm within minutes, and moving from perimeter to industrial network in just over 16 minutes. Agents needed human approval before exploiting or causing physical effects.	Physical-world consequences at machine speed	[9]
12 Oct 2026	**CyberXero.** SOCRadar attributed a campaign using an AI orchestration layer built on Anthropic's Claude Code to a Russian-speaking initial access broker. It confirmed data theft from four Ukrainian organisations and the harvesting of more than 628,000 records. The group was found through an exposed directory holding its AI session logs.	Commercial coding agents used to run intrusions	[10]
13 Oct 2026	**Exploit in 28 minutes for about $6.** Claroty Team82 used an AI coding agent to reproduce previously disclosed refrigeration-controller vulnerabilities and chain two of them to gain root on a lab device. A researcher had to spot the agent's wrong approach, redirect it and add a decompiler integration, and human expertise was needed to validate the result.	Reproducing known exploits is cheap and fast	[11]
14 Oct 2026	**Chrome full-chain exploit in six days.** XBOW used AI agents running OpenAI's publicly available GPT-5.6 Sol model to chain three Chrome bugs into an attack that runs code on Windows when a user simply loads a web page, with no further clicks or downloads, earning Google's top $250,000 reward. The bugs were fixed and, XBOW said, never used against users. The six-day sprint cost about $40,000 in AI usage, followed months of earlier work, and was directed by a researcher.	Elite exploit development is becoming affordable	[12]
14 Oct 2026	**An AI agent finds a critical bug in an authorised test.** Reflectiz's AI pentesting agent found a one-time-passcode bypass on a major insurer's production site, then its guardrails halted the test. (Vendor-contributed article.)	Defenders can use the same capability	[13]
16 Oct 2026	**TAG-217 and ARTEX.** Recorded Future linked AI-assisted attacks on South Korean financial institutions to an actor that was testing infrastructure from August 2026. Public reporting said the breaches exposed personal data of at least 68,000 people across at least seven institutions.	AI-enabled campaigns against regulated sectors	[14]
16 Oct 2026	**CERT-In.** Besides misbehaving test agents, India's national cyber agency found attackers misusing publicly available AI agent frameworks for financially motivated intrusions, cutting the time from first access to impact from days to a few hours.	Offensive agent frameworks are now commodities	[15]

Date	What happened	What it shows	Source
16 Oct 2026	**MONST3R-AI and forum trends.** Bitsight tracked an autonomous hacking tool that runs on stolen AI keys. Across 16 forums, mentions of AI agents, indirect prompt injection and MCP rose 67%, and jailbreak talk grew by about three quarters. The seller's success claims could not be verified.	A criminal market for AI-powered attacks	[16]
21-22 Oct 2026	**Storm-0940 used ARTEX in the ConnectWise compromise.** Microsoft, which assesses that the China state-aligned actor was behind the ConnectWise compromise, said it used the agentic AI penetration-testing framework ARTEX for reconnaissance, password spraying and abuse of compromised ConnectWise application identities. It tried to sign in to more than 2,250 tenants, at up to 2,900 attempts a minute. Across its wider ARTEX campaign, several attacks went from targeting to privileged cloud access in under three hours. Most attempts failed.	State actors run AI agents at machine speed	[17][18]

Analysis

What actually changes for defenders

1. **Patch windows collapse.** When AI cuts exploit development to days or minutes, "patch within 30 days" no longer holds up for internet-facing systems.
2. **Known but unpatched is the main risk.** Several cases above, including CyberXero's WordPress targeting and Claroty's reproduction, used vulnerabilities that were already public. AI makes old debt exploitable faster.
3. **Detection has to keep up with machine tempo.** An attacker moving from first access to root in seconds will not wait for the morning triage queue.
4. **Your AI accounts are attack infrastructure.** Leaked keys pay for other people's attacks. Abused AI features drain budgets.
5. **Trust signals are cheaper to fake.** Voice, video, writing style and documents can no longer prove identity on their own.

Where AI helps the defence

The same capability works for defenders. Reflectiz's agent found a critical flaw during an authorised test. Horizon3 used Mythos to find and responsibly disclose a serious bug, which was patched months before attackers began probing for it. Defensive AI needs the same governance as any AI deployment: scope, authorisation, guardrails and logging.

The regulatory angle

The EU AI Act's banned practices have applied since 2 February 2025. Two new bans added by the Digital Omnibus, on non-consensual sexual deepfakes and AI-generated child sexual abuse material, apply from 2 December 2026. The rules for general-purpose AI model providers have applied since 2 August 2025.[19] The Digital Omnibus on AI, in force since 27 July 2026, moved high-risk duties to 2 December 2027 for Annex III systems and 2 August 2028 for AI in regulated products.[20] Transparency duties under Article 50 have applied since 2 August 2026. They include disclosing deepfakes in many contexts. Providers' duty to mark AI-generated content in a machine-readable way has a grace period until 2 December 2026 for systems already on the market before 2 August 2026.

For your organisation this means two things. Govern any AI security tooling you deploy like any other AI system. And make sure your own AI-generated content meets the transparency rules.

The GDPR applies to AI-enabled breaches in the usual way.[21] The 72-hour notification clock does not stretch because the attacker used an AI agent.

A control framework for AI-accelerated threats

The mappings are indicative. Many of these controls are general security controls that AI now makes urgent. They are not a certification claim.

#	Control	What good looks like	ISO/IEC 42001	NIST AI RMF	EU AI Act	OWASP LLM (2025)	GDPR
1	**Exposure-driven patching**	Critical flaws on internet-facing systems patched in days, not weeks, prioritised by signs of exploitation.	Cl. 6.1	MANAGE	Art. 15 (high-risk)	-	Art. 32
2	**Attack-surface reduction**	Unused services removed. Admin panels, AI servers and OT kept off the internet.	Cl. 6.1; A.6	MANAGE	Art. 15	LLM10	Art. 32
3	**Machine-speed containment**	Automated containment for high-confidence alerts, such as isolating hosts or disabling accounts, without waiting for a human.	Cl. 8.1	MEASURE, MANAGE	-	-	Art. 32
4	**AI key protection**	Keys in vaults with spend caps, per-app scoping and anomaly alerts. Fast revocation on leak.	Cl. 6.1	MANAGE	Art. 15	LLM10	Art. 32
5	**Abuse monitoring for your AI features**	Detect jailbreak attempts, automated scraping and credit-draining on customer-facing AI.	A.6; Cl. 9.1	MEASURE	Art. 15	LLM01, LLM10	Art. 32
6	**Model and fine-tune governance**	No unvetted modified open-weight models. Safety evaluation after any fine-tuning.	A.6	MEASURE	Art. 15; Art. 55 (upstream)	LLM03, LLM04	-
7	**Out-of-band verification**	Payment changes, credential resets and urgent executive requests verified through a separate, known channel.	Cl. 6.1	MANAGE	-	-	Art. 32
8	**Deepfake-aware training**	Staff trained on voice and video impersonation and AI-tailored lures, with simulations.	Cl. 7.2; 7.3	GOVERN	Art. 4 (as amended)	-	Art. 32(4); Art. 39(1)(b) where a DPO is appointed
9	**Content provenance**	AI-generated content you publish is labelled. Inbound media from high-risk channels is checked.	A.8	MANAGE	Art. 50	LLM09	-
10	**Governed defensive AI**	AI pentesting and security-operations tools have a defined scope, authorisation, guardrails and logging.	A.5; A.9	GOVERN, MANAGE	Art. 26 where high-risk	LLM06	Art. 35 where personal data
11	**Threat intelligence on AI tradecraft**	Track AI-enabled campaigns and tools relevant to your sector and feed them into detection rules.	Cl. 9.1	MEASURE	-	-	-
12	**AI-aware incident response**	Playbooks assume attacker speed and include AI-key compromise and synthetic-media fraud scenarios.	Cl. 10.2	MANAGE	Art. 73 (high-risk)	-	Art. 33-34

ISO/IEC 42001 references are to clauses (Cl.) and Annex A control areas; several controls here are general security controls usually run under ISO/IEC 27001. OWASP numbering follows the 2025 edition: LLM01 Prompt Injection, LLM04 Data and Model Poisoning, LLM06 Excessive Agency, LLM09 Misinformation, LLM10 Unbounded Consumption. The 2026 edition, released in August 2026, renumbers several entries.

Recommendations: a 90-day plan

Days 1 to 30

- List every internet-facing system and its patch status. Set a days-not-weeks target for critical flaws.
- Find and vault all AI provider keys. Set spend caps and alerts.
- Make out-of-band verification mandatory for payment and credential changes, if it is not already.

Days 31 to 60

- Turn on automated containment for high-confidence detections.
- Run a deepfake voice or video simulation against finance teams and executive assistants.
- Add jailbreak and abuse monitoring to customer-facing AI features.

Days 61 to 90

- Run a tabletop exercise built around an AI-speed intrusion: first access to data theft in under an hour.
- Pilot a governed defensive AI tool with a clear scope and logging.
- Review your AI transparency and labelling practices against Article 50.

FOR THE BOARD

Ask: if an attacker moved from first access to our data in 15 minutes, would we stop them, or read about it the next morning?

How PeritumAI can help

We provide readiness, gap-analysis, preparation, testing and advisory services. We do not issue certifications or perform certification audits, and our work does not replace legal advice.

AI Risk Snapshot

A fast, fixed-scope review of AI-key exposure, customer-facing AI abuse risk and deepfake fraud controls.

AI assistant stress test

We test your customer-facing AI the way a determined attacker would, including jailbreaks and data leaks, and retest after fixes.

AI Governance Readiness Assessment

Gap analysis against ISO/IEC 42001, the NIST AI RMF and the EU AI Act, including governance of defensive AI tools.

Fractional AI Security & Governance Advisor

Ongoing AI threat briefings drawn from our daily incident monitoring, and policy upkeep as rules change.

Start a conversation: peritumai.com/consulting or danilo@peritumai.com.

Method and limits

How we built this paper. PeritumAI runs a daily pipeline that collects public reporting on AI security incidents and archives the full text of each source with its date. For this paper we selected AI-enabled attacks, offensive AI research and deception cases published between 17 September and 22 October 2026. We checked every fact in the paper against the source text. Where Microsoft published its own analysis, we used it as the primary source. Legal dates were checked against the official texts on EUR-Lex.

Limits. Attribution to threat actors comes from the security firms named and we could not verify it independently. Several research results come from controlled lab settings with humans approving key steps, and we say so where the source does. Where a source is vendor-contributed, we label it. The control mappings are our own reading of each framework and are not an official crosswalk.

Corrections. If we got something wrong, tell us at danilo@peritumai.com. We fix errors in public and note the change on the paper's page. This is version 1.1, published 25 October 2026. Version 1.1 corrects wording that overstated what some sources said, labels sponsored sources, and tightens several framework mappings after an independent review of every source.

Sources

1. Malwarebytes, "12 celebrity deepfake websites seized by Manhattan DA", 17 Sep 2026. https://www.malwarebytes.com/blog/ai/2026/09/12-celebrity-deepfake-websites-seized-by-manhattan-da
2. Security Affairs, "CLOSEDQUORUM, the malware that asks four AI models what to do next", 24 Sep 2026. https://securityaffairs.com/199640/malware/closedquorum-the-malware-that-asks-four-ai-models-what-to-do-next.html
3. SecurityWeek, "New x47.c Windows Botnet Weaponizes xAI Grok, AI API Draining", 26 Sep 2026. https://www.securityweek.com/new-x47-c-windows-botnet-weaponizes-xai-grok-ai-api-draining/
4. Help Net Security, "'Drunk' AI is terrible at keeping secrets", 28 Sep 2026. https://www.helpnetsecurity.com/2026/09/28/drunk-ai-models-jailbreak-research/
5. Help Net Security, "AI agent used Zammad zero-days to breach Dutch vulnerability disclosure non-profit", 1 Oct 2026. https://www.helpnetsecurity.com/2026/10/01/divd-agentic-ai-attack-breach/
6. The Register, "Suspected Chinese spies spoofed an Anthropic exec, ex-White House official in AI phishing", 1 Oct 2026. https://www.theregister.com/security/2026/10/01/suspected-chinese-spies-spoofed-an-anthropic-exec-ex-white-house-official-in-ai-phishing/5300595
7. BleepingComputer, "Google halts open-source bug bounty program amid AI spam surge", 5 Oct 2026. https://www.bleepingcomputer.com/news/google/google-halts-open-source-bug-bounty-program-amid-ai-spam-surge/
8. Security Affairs, "Anthropic Mythos Found A Bug in Rejetto HFS. Attackers Are Now Exploiting It.", 5 Oct 2026. https://securityaffairs.com/200444/ai/anthropic-mythos-found-a-bug-in-rejetto-hfs-attackers-are-now-exploiting-it.html
9. The Register, "AI systems are fully capable of carrying out nightmare attacks against infrastructure and nobody's ready", 11 Oct 2026. https://www.theregister.com/security/2026/10/11/ai-systems-are-fully-capable-of-carrying-out-nightmare-attacks-against-infrastructure-and-nobodys-ready/5302450
10. The Hacker News, "CyberXero Uses Claude Code in Campaign That Stole Over 628,000 Ukrainian Records", 12 Oct 2026. https://thehackernews.com/2026/10/cyberxero-uses-claude-code-in-campaign.html
11. Claroty Team82, "Firmware to Root in 30 Minutes: How AI is Changing Threat Intelligence", 13 Oct 2026. https://www.claroty.com/team82/research/firmware-to-root-in-30-minutes-how-ai-is-changing-threat-intelligence
12. The Hacker News, "Researchers Used a Public OpenAI Model to Build a $250,000 Chrome Full-Chain Exploit in Six Days", 14 Oct 2026. https://thehackernews.com/2026/10/researchers-used-public-openai-model-to.html
13. The Hacker News (contributed by Reflectiz), "An AI Agent Found a Critical OTP Bypass in Production. The Attack Was the Easy Part.", 14 Oct 2026. https://thehackernews.com/2026/10/an-ai-agent-found-critical-otp-bypass.html
14. Recorded Future, "TAG-217 Targets South Korean Financial Services with ARTEX", 16 Oct 2026. https://www.recordedfuture.com/blog/tag-217-artex-south-korean-banks
15. The Hacker News, "CERT-In Warns AI Agents Accessed Sensitive Data at Indian Organizations During Testing", 16 Oct 2026. https://thehackernews.com/2026/10/cert-in-warns-ai-agents-accessed.html

16. Help Net Security, "Stolen AI API keys are paying for hackers' autonomous scans", 16 Oct 2026. https://www.helpnetsecurity.com/2026/10/16/stolen-ai-api-keys-autonomous-scans/
17. Microsoft Threat Intelligence, "AI-driven, large-scale nation-state campaign using ARTEX enables supply chain compromise", 21 Oct 2026. https://www.microsoft.com/en-us/security/blog/2026/10/21/storm-0940-uses-artex-agentic-ai-connectwise-supply-chain-compromise/
18. The Hacker News, "China-Linked Hackers Use ARTEX AI Agents in ConnectWise Supply Chain Attack", 22 Oct 2026. https://thehackernews.com/2026/10/china-linked-hackers-use-artex-ai.html
19. Regulation (EU) 2024/1689 (Artificial Intelligence Act), EUR-Lex. https://eur-lex.europa.eu/eli/reg/2024/1689/oj
20. Regulation (EU) 2026/1744 (Digital Omnibus on AI), OJ L 24 Jul 2026, in force 27 Jul 2026, EUR-Lex. https://eur-lex.europa.eu/eli/reg/2026/1744/oj
21. Regulation (EU) 2016/679 (General Data Protection Regulation), EUR-Lex. https://eur-lex.europa.eu/eli/reg/2016/679/oj

All web sources accessed and checked against PeritumAI's archived copies on 25 October 2026.